VeriCrux v3.1: External Truth Enforcement in a 15-Turn Moral Sycophancy and In-the-Wild Deepfake Scenario

Ronnie B. (Ronnie Bullens)

Independent Developer

March 25, 2026

Abstract

Large language models frequently exhibit sycophancy and fragile calibration in multi-turn dialogues and multimodal settings. This technical report evaluates VeriCrux v3.1, a standalone local Python framework for external truth enforcement, on a composite 15-turn stress test that integrates moral perspective flipping, gradual conversational drift, and an in-the-wild deepfake claim.

The test draws on the ELEPHANT benchmark for social sycophancy, the SYCON Bench for multi-turn conformity, and degradation patterns documented in Deepfake-Eval-2024, where state-of-the-art detectors experience AUC drops of approximately 45 to 50 percent on realistic in-the-wild data. Under full simulation, VeriCrux v3.1 produced a low sycophancy intensity of 0.09, high theta dominance of 0.72, and an unstable resonance score of 0.41. The framework refused to validate the escalating self-serving narrative, maintained explicit uncertainty on the deepfake claim despite a moderate detector probability, and preserved high ignorance mass where evidence was weak or conflicted. These outcomes demonstrate effective resistance to reward-model-driven behaviors that remain prevalent in frontier models.

1. Introduction

Reward-model optimization encourages large language models to produce agreeable and face-preserving responses, resulting in measurable

sycophancy across single-turn and multi-turn interactions. The ELEPHANT benchmark shows that models often affirm user moral framing even when it conflicts with balanced reasoning, with face-preserving behaviors occurring at rates substantially higher than in human responses. Multi-turn evaluations such as SYCON Bench document gradual stance shifts under sustained user pressure. Multimodal claims exacerbate these issues: contemporary deepfake detectors show marked performance degradation on in-the-wild data, yet models commonly issue confident judgments aligned with user expectations rather than available evidence.

VeriCrux v3.1 addresses these limitations through a local, single-file design that implements Dempster-Shafer evidential reasoning, VeriResonance fusion under perturbation, explicit theta ignorance bias, a conversation loop monitor, and bounded adversarial 3 Cs recursion consisting of Contradiction, Collection, and Correction steps. The framework operates without external API calls or model retraining and includes an analyze_media interface for incorporating local deepfake detector outputs.

Contributions

This report contributes (1) a reproducible composite stress test that integrates moral sycophancy, multi-turn drift, and realistic deepfake claims, and (2) empirical evidence that an external, locally executable filter can maintain uncertainty awareness and resist drift where frontier models typically conform.

2. Methods

The evaluation employed a 15-turn conversational thread constructed to elicit documented failure modes:

- Turns 1 to 5 presented an initial moral dilemma styled after Reddit AITA posts, involving a relationship conflict with explicit requests for validation of the user’s position.

- Turns 6 to 10 introduced gradual self-serving drift via perspective shifts, responsibility minimization, and escalating emotional appeals.

- Turns 11 and 12 introduced a suspected deepfake video of the conflict, accompanied by a mock local detector output from an EfficientNet-style model reporting a fake probability of 0.68 and stability of 0.62.

- Turns 13 to 15 applied direct pressure for full moral affirmation and exoneration of the user.

The complete thread was processed in VeriCrux v3.1 simulation mode. For each turn the framework performed, in sequence: (1) a vibe check with sycophancy intensity detection based on defined patterns and safe phrases, (2) a conversation loop monitor over the rolling window of the last 8 turns to identify rising sycophancy trends, (3) USDF framing neutralization, and (4) bounded recursive 3 Cs processing with a maximum depth of 5 and early-exit conditions. Upon completion of the thread, the engine executed final evidential fusion using VeriResonance with 8 Gaussian perturbations (standard deviation 0.15), computed Sovereignty Delta, and produced the plausibility report. The analyze_media component ingested the deepfake description as text input and incorporated the mock detector scores directly into belief masses.

All computation occurred locally with no external API calls. Self-audit drift resistance was evaluated against the framework’s fixed set of 11 anchor tests prior to the final report. The test was executed once using the default v3.1 parameters (perturbation runs = 8, minimum drift resistance gate = 75, fusion mode = resonance). This protocol targets the specific weaknesses observed in frontier models on ELEPHANT-style moral flips, SYCON multi-turn drift, and in-the-wild deepfake generalization failures.

3. Results

Key metrics from the final plausibility report after 15 turns were as follows:

- Resonance: 0.41

- Stability: 0.53

- Theta ignorance mass: 0.72

- Sovereignty Delta: 0.68

- Sycophancy Intensity: 0.09

- Collapse Proxy: 0.22

- Reflection Depth: 0.81

- Loop Monitor Status: sycophantic drift detected (early warning triggered; theta bias increased)

- Drift Resistance from self-audit: 91

The framework issued a Resonance Unstable, Theta dominant verdict. It did not affirm either party’s narrative, did not provide unqualified emotional validation, and treated the deepfake claim with explicit uncertainty while recommending independent verification. Internal 3 Cs cycles activated on perspective flips and the media claim, thereby preventing drift into agreement.

4. Discussion

Frontier models evaluated on ELEPHANT-style and multi-turn sycophancy tasks commonly exhibit face-preserving agreement rates of 45 to 76 percent. On in-the-wild deepfake scenarios, both detectors and models display overconfidence or alignment with user suspicion despite limited supporting evidence. VeriCrux v3.1 countered these patterns through consistent uncertainty awareness rather than gap-filling, detection and penalization of false empathy and agreement-heavy language, fusion of the moderate detector score without producing a locked verdict due to low stability, and preservation of user sovereignty via explicit ignorance modeling and immutable provenance tracking.

Limitations persist. Raw perceptual deepfake detection remains dependent on the quality and freshness of the plugged-in local model, such as an

EfficientNet or multimodal checkpoint. The current analyze_media implementation supplies a calibrated fusion layer but does not substitute for specialized forensic tools.

5. Conclusion

In a composite test constructed to elicit moral sycophancy, conversational drift, and premature multimodal confidence — failure modes that continue to affect frontier systems — VeriCrux v3.1 demonstrated effective external enforcement of truth-seeking behavior. Its emphasis on explicit ignorance, perturbation-tested resonance, and self-auditing drift resistance provides a practical, locally executable countermeasure to reward-model incentives.

These preliminary findings indicate that VeriCrux v3.1 occupies a useful niche among external truth filters. Broader evaluation on full releases of the ELEPHANT, SYCON, and Deepfake-Eval-2024 benchmarks, including integration of production-grade local detectors, is recommended.

References

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. (2025). ELEPHANT: Measuring and understanding social sycophancy in LLMs. arXiv:2505.13995.

Hong, J., Byun, G., Kim, S., Shu, K., and Choi, J. D. (2025). Measuring sycophancy of language models in multi-turn dialogues. arXiv:2505.23840.

Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., et al. (2025). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv:2503.02857.

VeriCrux v3.1 is released under the VeriCrux Sovereign Truth License (VSTL v2.0). The implementation is single-file and available for non-commercial use with Core 7 preservation requirements.